TATION PAGE

Form Approved
OMB No 0704-0188

2

tc average 1 hour per response including the time for reviewing instructions searching existing data sources wing the collection of information Send comments regarding this burden estimate or any other aspect of this en to Washington Headquarters Services Directorate for Information Operations and Reports 1215 Jefferson e of Management and Budget Paperwork Reduction Project (0704-0188) Washington DC 20503

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
June 1990

3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE

Graphical displays of: Are the (x, y) pairs compatible with a linear dependence?

5. FUNDING NUMBERS
DAAL03-88-K-0045

6. AUTHOR(S)

John W. Tukey

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 26098.22-MA-SDI

11. SUPPLEMENTARY NOTES

The view, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

A special case of importance in the display of simultaneous intervals (SCL's or SPL's) arises when the results of measurement or observations on y at corresponding values of x are to be displayed with the intent of understanding what they say about (a) the plausibility of the existence of a smooth dependence on x of the underlying value ave$\{y \mid x\}$ (the average that would have been found were it possible to repeat the measurement y very many times for the same x), and (b) which smooth dependencies seem to be plausible in view of the data.

The display techniques focusing on matched SCL's and SPL's developed in Technical Report No. 300 [Tukey 1990] are not particularly useful here. The aperture (pencil-point") techniques of Hoaglin and Tukey (1985n) suggest very useful approaches, although, as reported in that reference, these techniques have so far been directed toward the combination of SCL's and ICL's. The present account builds on all the insights thus developed, seeking for simplicity and treating related problems as they arise.

14. SUBJECT TERMS

Apertures, Compatibility, Confidence limits, Graphical display, Partial limits, SCL's, SPL's, SQL's, Saskatchewan data, Severe limits, Significance of lack of fit, Simultaneous limits, Tight limits.

15. NUMBER OF PAGES
27

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Graphical displays of: Are the (x,y) pairs compatible with a linear dependence?

John W. Tukey
Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000

Technical Report No. 301
Department of Statistics
Princeton University
Princeton, NJ 08544-1000

June 1990

Prepared in connection with research at Princeton University sponsored by the Army Research Office (Durham), DAAL03-88-K-0045.

TABLE OF CONTENTS

LIST OF EXHIBITS

Graphical displays of: Are the (x, y) pairs compatible with a linear dependence?

John W. Tukey

Technical Report No. 301

Princeton University
Fine Hall
Washington Road
Princeton, NJ 08544-1000

KEYWORDS: Apertures, Compatibility, Confidence limits, Graphical display, Partial limits, SCL's, SPL's, SQL's, Saskatchewan data, Severe limits, Significance of lack of fit, Simultaneous limits, Tight limits.

1. Introduction

A special case of importance in the display of simultaneous intervals (SCL's or SPL's) arises when the results of measurem^nt or observations on y at corresponding values of x are to be displayed with the intent of understanding what they say about (a) the plausibility of the existence of a smooth dependence on x of the underlying value ave$\{y \mid x\}$ (the average that would have been found were it possible to repeat the measurement y very many times for the same x), and (b) which smooth dependencies seem to be plausible in view of the data.

The display techniques focusing on matched SCL's and SPL's developed in Technical Report No. 300 [Tukey 1990] are not particularly useful here. The aperture (pencil-point") techniques of Hoaglin and Tukey (1985n) suggest very useful approaches, although, as reported in that reference, these techniques have so far been directed toward the combination of SCL's and ICL's. The present account builds on all the insights thus developed, seeking for simplicity and treating related problems as they arise.

2. Intervals and apertures

We assume that our y's come with a useful estimate of their uncertainty, useful in the sense of providing (approximate, of course) probability statements about ranges of plausible values for what they measure. Thus we might want to state an ICL (individual confidence interval) for each point, for example that, we have

$$95\% \text{ confidence that } 11.2 \le \text{ave}\{y \mid x_i\} \le 12.8$$

How should we display such a statement graphically?

The classical approach is to picture one interval from 11.2 to 12.8 with a line segment or bar, perhaps decorated with arrowheads, etc. This stresses what we do *not* know by stressing the variety of possible values that are plausible. This is perhaps the natural first step beyond

Prepared in connection with research at Princeton University sponsored by the Army Research Office (Durham), DAAL03-88-K-0045.

June 13, 1990

the most naive picture - - in our example a dot at or near the midpoint of the ICL, 12.0. But it is not a good way to go on to more complex situations.

A better approach for many purposes is to plot what we know rather than what we do not know - - in our example to draw something from a low value to 11.2 and another something from 12.8 to a high value; thus "blackening" the implausible values for $\text{ave}\{y \mid x_0\}$ and showing our range of uncertainty by an aperture - - a "hole in the fence"! A form of this, using "pencil point" to combine ICL's and SCL's, was proposed by Hoaglin and Tukey (1985n) We deal here with a similar problem but in a different context. So we are led to somewhat different graphical display that, however, use the same approach - - apertures.

3. Classes of questions

If we have a set of (x, y) pairs, which might reasonably represent more or less random fluctuations around a systematic dependence, there is a natural set of pairs of questions that can be asked about the supposed dependence, each pair of the form:

- Is it reasonable that there is a dependence of a specified class?
- If so, what subclass of dependencies of this class is reasonable?

The classes for which we are most likely to ask these pairs of questions are, in order, from simple to complicated:

1) *constant dependence* ($\text{ave}\{y \mid x\} = C_0$ for all x, and some unknown C_0)

2) linear dependence ($\text{ave}\{y \mid x\} = C_0 + C_1 x$, for all x, and some unknown C_0, C_1)

3) some special dependence (according to subject-matter field)

4) monotone dependence (either $\text{ave}\{y \mid x_1\} \le \text{ave}\{y \mid x_2\}$ whenever $x_1 \le x_2$, or $\text{ave}\{y \mid x_1\} \ge \text{ave}\{y \mid x_2\}$ whenever $x_1 \le x_2$)

5) dependence monotone, except for one maximum *or* one minimum (sense of monotonicity changes at that extremum).

It is much easier to continue this list than to deal with the probability problems that (4) or (5) suggest. So we leave extensions to the reader.

4. The probabilistic situation

An important consideration is the difference between

A) Is this *specific* dependence of the chosen class plausible?

and

B) Is *any* dependence of the chosen class plausible?

So far as general situations go, the naive way to answer (B) amounts to asking (A), in parallel, for each possible specific dependence of the class, and announcing whether any of them is plausible. Quite naive, but not best.

In the simplest case, where we ask if ave$\{y \mid x\}$ might be the same for all x, if we represent each measurement or observation by an aperture, Technical Report No. 300 (Tukey 1990), made a clear distinction between (tight or severe) SCL's and SPL's. Here SCL's are simultaneous confidence limits, which, in our present style, would surround each data point with an aperture - - an SCA - - such that a value of C is compatible with that measurement if the horizontal line at height C passes through the aperture. SPL's, on the other hand, are simultaneous partial limits which, in our present style would correspond to narrower apertures - - SPA's - - such that two y's might have the same underlying value if there is some horizontal line that passes through *both* apertures.

For horizontal lines, the best answer - - or class of answers - - to (B) is thus provided by tight simultaneous partial apertures (SPA's), just as the best answer - - or class of answers - - to (A) for horizontal lines is provided by tight simultaneous confidence apertures (SCA's).

* slanting lines *

Consider next the case of slanting lines, where

$$\text{ave}\{y \mid x\} = C_0 + C_1(x - x_{00})$$

If we knew the slope, C_1, we could replace y by $y - C_1(x - x_{00})$, reducing the problem to the case of horizontal lines, just discussed. Thus the answers to

A) Is this *specific* straight-line dependence (*with slope* c_1) *plausible?*

and

B) Is *any* straight line dependence (with slope c_1) plausible?

are best given, by passage or non-passage of a line of the given slope, respectively, through (A) simultaneous confidence apertures or (B) simultaneous *partial* apertures.

If, as is usually the case, we do *not* know the slope, we cannot confine our attention to any one slope. Thus the chance that a line will pass *all* simultaneous partial apertures, when all the values of $(x, \text{ave}\{y \mid x\})$ lie on some unknown line, will be larger than SPL nominal. This will happen because (a) the chance that a line of correct slope will pass equals the nominal and (b) it is possible that, while *no* line of correct slope passes, one of incorrect slope *does* pass.

Since Q follows P in the alphabet, let us define tight SQL's and SQA's to be the limits or apertures such that the chance, in the null situation, of a line passing all of them is the nominal level. Then

$$\text{coefficient of tight SQL} < \text{coefficient of tight SPL}$$

and

(tight) SQ-apertures are only part of (tight) SP-apertures

For general k - - and for general patterns for the k x's - - there seems to be no available tabulation (or useful closed form expression) for the tight SQL coefficients. (If we had answers for k equally spaced x's, however, we would probably be in relatively good shape.) For k=3 and equally spaced x's, however, it is easy to calculate the SQL coefficients. If we do this (see Section 10) and go on to ask what approximation is suggested for more general k, we are led to believe that

tight SPL coefficient – .68 (severe SCL coefficient – severe SPL coefficient)

is probably not a bad approximation to

tight SQL coefficient.

(See Section 11 for a different approximation.

We return below, in Section 12, to the question of severe SQL coefficients, which seem to have quite limited utility.

5. Display choices

In Technical Report 300, we argued the advantages of a horizontal bar in displaying SP intervals and a point in displaying SC intervals. This led to a double open triangle arrowhead style of display. If we convert this directly from intervals to apertures, we move from the style of the left-hand side of exhibit 1 to the style of that exhibit's center. The results are relatively effective so long as only horizontal lines are candidates to pass through the apertures. (Since this is the case where SP apertures are appropriate, this is not a serious restriction.)

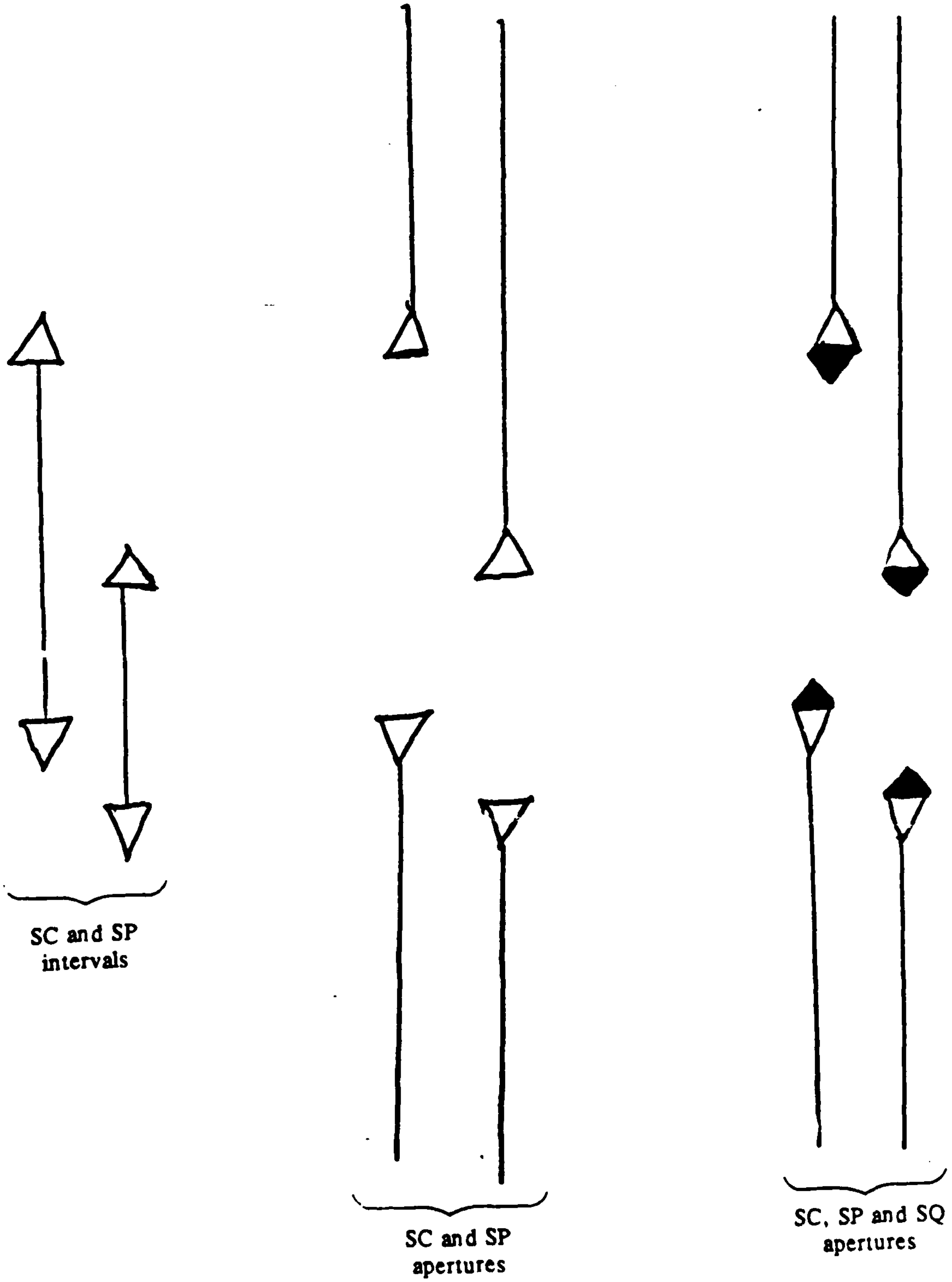
exhibit 1
about here

When we want to go to SQ apertures, plausibly by adding to what is already displayed, we want:

- a pointed glyph, since lines need avoid the aperture only at its own value of x,
- a reasonably emphatic glyph,
- a glyph that directs our attention inward.

exhibit 1

Evolution of intervals to apertures



The style of the right-hand side of exhibit 1 meets all three of these desiderata, while preserving the double open triangles we chose earlier for SC and SP.

In the half-open quadrilaterals (hoquas) of the new style:

- the black tip marks the end of the SQA
- the black-white division marks the end of the SPA
- the white tip marks the end of the SCA.

We shall use this style until a better one is found, using our roughly approximate SQA's until better coefficients are availabl:. When, as, and if, such coefficients are available for slanting lines we expect to use them instead. When the class of candidates is broader than all straight lines, we will want to use even narrower apertures. The time when we will have corresponding coefficients for each of several classes of coefficients seems uncertain. The best temporary solution seems to be to use the best available SQ coefficients, and to continue to carry along a substantial grain of salt.

6. The Saskatchewan data

Brillinger (1990) has recently illustrated a moderately complicated analysis of geographically aggregated data, using 1986 births in the 18 Census divisions of Saskatchewan, as an example. We will use the same data as an example of a simpler analysis using SQ apertures.

* the data *

The basic population counts, by division, for the division as a whole, and for the 3 largest places in the division (of size 1,000 or more) is set out in exhibit 2. The last two columns show the ratios of (a) population of largest place and (b) sum of populations of 3 largest places to the total population. We will return to these and related figures shortly.

exhibit 2
about here

The minimum instability we can plausibly apply to a number of births - - unless we take a narrowly historical view - - is that of the Poisson distribution. To more accuracy than we are likely to need in the present example, a Poisson distribution corresponds to

$$\text{score} = \sqrt{2 + 4(\text{count})}$$

following a Gaussian distribution with unit variance.

Thus, given a coefficient h for some form of simult:neous aperture, we can locate the ends of the aperture by applying the inverse transformation

exhibit 2

Populations in the 18 census divisions
(3K57 means "3 thousand, 570", etc.)

Division	Largest	Next	Third	Total	L/T	3/T
1	9K17 Estevan	1K19 Oxbow	1K07 Carlyle	32,771	.28	.28
2	9K52 Weyburn	1K01 Radville	?	26,677	.36	.43
3	2K92 Assiniboine	1K34 Gravelbourg	1K03 Coronach	26,367	.14	.26
4	2K47 Maple Creek	2K11 Shaunavon	?	14,659	.17	.37
5	5K09 Melville	3K06 Esterhazy	2K58 Moosomin	41,202	.12	.26
6	162K61 Regina	1K89 Indian Head	1K83 Fort Qu'Appelle	201,021	.81	.83
7	33K94 Moosejaw	1K02 Herbert	?	52,875	.64	.68
8	14K75 Swift Current	1K41 Eston	1K11 Leader	35,146	.42	.49
9	15K15 Yorkton	2K69 Kamsack	2K67 Canora	44,918	.34	.46
10	2K15 Wynyard	1K49 Wadena	1K45 Foam Lake	25,121	.09	.20
11	154K2 Saskatoon	2K08 Warman	1K97 Outlook	192,551	.80	.82
12	3K56 Battleford	2K66 Rosetown	2K56 Biggar	25,493	.14	.34
13	3K97 Kindersley	2K41 Unity	1K50 Wilkie	27,375	.14	.28
14	6K01 Melfort	4K38 Nipawin	3K11 Tisdale	47,467	.13	.28
15	31K38 Prince Albert	4K7 Humboldt	1K61 Rosthern	79,980	.39	.47
16	14K03 N. Battleford	1K23 Shellbrook	?	39,905	.35	.40
17	6K03 Lloyd Minster	3K86 Meadow Lake	1K00 Maidstone	35,481	.17	.31
18	1K64 Creighton	?	?	25,304	.06	.14

L/T = ratio of largest to total

3/T = ratio of sum of three largest with (?) filled in as 0K90 = 900)

$$\text{corresponding count} \approx ((\text{score})^2 - 2)/4$$

to the ends of the aperture

$$\text{observed score} \pm h$$

in terms of the score. To get ends of an aperture for birth rate we need to divide the corresponding counts by the number of women of appropriate age (Brillinger used the 25-29 age group).

Turning back to Technical Report 300, with $k=18$ and $\nu=\infty$ (under the Poisson assumption, the variance is known!), we find coefficients of ±2.98 for severe SC, ±2.53 for severe SP and ±2.46 for tight SP. Our working approximation now gives ±2.16 for SQ. Accordingly, exhibit 3 gives numbers of women and births (both of which have been modified, see Brillinger 1990 for details, for privacy reasons) overall birth rates, and the 3 sets of birth rate intervals.

exhibit 3
about here

* urbanicity vs. rurality *

As Brillinger noted, birth rates tended to be low in more urban divisions, high in more rural ones. If we are to picture what we know with a well-placed set of apertures, we need an x-variable that reflects rurality vs. urbanness. Kafadar and Tukey (1991, and papers in preparation) have faced this question with regard to county-by-county death rates from various forms of cancer. Their conclusions include:

- steps of 1/2 in the logarithm to the base 2 of the size of the largest place in the county seems to do quite well (Kafadar and Tukey 1991);
- especially in certain suburban areas, we do well to allow the sizes of the 2nd and 3rd largest places to influence the classification;
- working with the square-root of the sums of squares of the 3 largest places seems one reasonable approach.

Thus it seemed plausible to approach the Saskatchewan data in a somewhat similar way.

* Zipfing the tail *

The most easily available place-by-place population data is very likely to cut off at some size (perhaps 2,500 or 1,000). In Saskatchewan, even with a cut off at 1,000, this left Census divisions with less than 3 places of recorded population. Thus it was natural to ask whether there was any easy way to approximate the missing populations.

exhibit 3

Observed birth rates and intervals based
on ±2.98, ±2.46, and ±2.14 applied to $(2 + 4(\text{births}))^{1/2}$
(irrelevant decimal places given)

Division	birth rate	±2.98		±2.46		±2.14*	
1	.169	.1359	.2034	.1412	.1970	.1446	.1933
2	.137	.1012	.1714	.1086	.1646	.1100	.1604
3	.142	.1082	.1924	.1146	.1841	.1186	.1791
4	.155	.1146	.2171	.1203	.2089	.1271	.2007
5	.164	.1320	.1996	.1374	.1931	.1407	.1892
6	.133	.1226	.1440	.1244	.1420	.1255	.1409
7	.139	.1149	.1614	.1187	.1570	.1210	.1544
8	.155	.1274	.1909	.1324	.1819	.1356	.1812
9	.158	.1262	.1908	.1313	.1846	.1345	.1809
10	.143	.1082	.1924	.1146	.1841	.1186	.1791
11	.131	.1214	.1413	.1231	.1395	.1241	.1389
12	.185	.1436	.2215	.1500	.2167	.1539	.2120
13	.167	.1334	.2037	.1389	.1970	.1424	.1929
14	.162	.1335	.1951	.1384	.1893	.1415	.1857
15	.148	.1270	.1673	.1303	.1636	.1324	.1613
16	.137	.1112	.1684	.1440	.2047	.1634	.2416
17	.171	.1440	.2047	.1489	.1990	.1519	.1955
18	.201	.1634	.2426	.1697	.2351	.1736	.3305

*This value lies between the approximations of Sections 4 and 10 (±2.16) and Section 11 (±2.11) for the tight SQL coefficient. Using either ±2.11 or ±2.16 instead of ±2.14 would make a small change in this table and no visible change in the pictures derived from it (for example, ±2.16 is only 1/16 of the way from 2.14 to 2.46!).

Many extreme J-shaped distributions, like those of populations of places in some local area, show behavior analogous to the rank-size rule of (Zipf's Law) according to which

(size)(rank from above) = constant

so that

$$\text{size of } g^{th} \text{ largest place} \approx \frac{\text{constant}}{g}.$$

(Systematic deviations, especially for small g or large g, are probably more common than close agreement. For some instances, see Tukey 1977, Chapter 18.)

This approximation *corresponds to simple ratios for* the sum of squares of sizes of all places smaller than a given place as a multiple of the squared size of that place and its rank g. Exhibit 4 shows the the results for small g.

exhibit 4
about here

In using this approximation, we need to take account of any cut-off on the list. Some examples from exhibit 2 will illustrate the opportunities. Division 2's second-sized place is close to 1,000, so that we can enter exhibit 4 with $g = 2$ to get an approximate sum of squares of sizes of all places from the third largest onward. Division 4 has a second-sized place above 2,000 and we know the third-sized must be below 1,000. Therefore we do better to choose a third-sized size and then turn to exhibit 4 with g=3. The largest choice for the 3rd-sized place is just below 1,000. We have chosen 1,000 in such cases, preferring to overestimate the remaining sum of squares somewhat.

Exhibit 5 shows, for each of the 18 divisions, the square roots of the sums of squares of sizes

- for the largest place
- for up to 3 places of size ≥ 1,000
- for all places, approximated as just described.

It also shows these sizes (found as square roots) as fractions of the total population.

exhibit 5
about here

One important aspect of this table is the similarity of the orderings provided by most columns both with each other and with those based on the last two columns of exhibit 2.

exhibit 4

Approximate ratios of sum of squares of sizes of all smaller places to the square of the size of the g^{th} largest place (based on rank-size rule)

g	ratio
1	.645
2	1.58
3	2.56
4	3.54
5	4.54
6	5.53
7	6.53
8	7.52
9	8.52

exhibit 5

Equivalent sizes of largest place for the 18 divisions
(201K0 is 201,0*xy*, etc., and irrelevant decimal places are carried to show similarity of ratios)

								(***)
Division	Total	Largest	(*)	(**)	L/T	(*)/T	(**)/T	16–2log$_2$(**)
6	201K0	162K61	162K64	164K66	.809	.809	.809	1.31
11	192K6	154K21	154K28	154K31	.801	.801	.801	1.46
7	52K87	33K94	33K96	33K98	.642	.642	.643	5.83
15	79K98	31K38	31K79	31K85	.392	.397	.398	6.01
9	44K92	15K34	15K45	16K02	.342	.353	.357	8.00
8	35K15	74K75	14K85	14K94	.420	.423	.425	8.20
16	39K90	14K03	14K08	14K17	.352	.353	.355	8.35
2	26K67	9K52	9K57	9K66	.357	.359	.362	9.46
1	32K77	9K17	9K37	9K57	.280	.286	.292	9.48
17	35K48	6K03	7K23	7K40	.170	.204	.209	10.22
14	47K47	6K01	8K54	8K84	.127	.180	.187	9.79
5	41K20	5K09	6K89	7K35	.124	.167	.178	10.24
13	27K37	3K97	4K87	5K33	.145	.178	.194	11.17
12	25K49	3K56	5K13	5K55	.140	.202	.210	11.05
3	20K36	2K92	3K46	3K83	.143	.170	.188	12.17
4	14K65	2K47	3K40	3K76	.168	.232	.256	12.16
10	25K12	2K15	2K99	3K67	.086	.119	.146	12.24
18	25K30	1K64	1K64	2K29	.065	.065	.091	13.61

(*) Square root of total squared population of *3 largest* places (of size ≥ 1000) in the division.

(**) Square root of approximate total squared population of *all* places in the division.

(***) Here (**) is expressed in thousands.

Notice that how far we go summing squares matters very little in the top half of this table.

Notice that ordering on "Largest" has produced a close approximation to order in each of the six columns to the right.

Notice that the difference between (*)/T and (**)/T never exceeds .027 and that the difference between 16–2log$_2$(*) and 16–2log$_2$(**) does not exceed 0.48, and for all but the four smallest divisions does not exceed 0.26.

Presumably each of these columns does a moderately reasonable job of displaying rurality-urbanicity. We shall build our analysis here on the last column of exhibit 5, but we encourage readers both to look at other choices and to cross-plot some of these columns against one another.

* re-expression *

If we plot birth rate as a response, and $16-2\ \log_2((**)/1000)$ as a circumstance, we get exhibit 6. Trend is clear, as is appreciable curvature.

exhibit 6
about here

A little trial urges us to use $(16-2\ \log_2((**)/1000))^2$ as the circumstance, and leads to exhibit 7, in which 5 divisions deviate to one side, while the other 13 lie reasonably well along a straight line.

exhibit 7
about here

7. Apertures in the example

If we now use the style suggested in Section 5 to display birth rate against $(16-2\ \log_2((**)/1000))^2$ we get the picture in exhibit 8. Clearly a reasonable variety of straight lines pass all apertures. At whatever approximation to 5% our approximate SQA's provide, then, the data are individually-but-collectively consistent with a linear relation of divisional birth rate to a simple measure of rurality.

exhibit 8
about here

It is clear that, as we would have anticipated, not all apertures contribute to restricting the set of piercing straight lines. It thus seems natural to produce a skeleton version of exhibit 8 which shows only the aperture edges that provide additional restriction, beyond that provided by others. Exhibit 9 provides this skeleton version, and includes the bounding positions of the lines that pierce all apertures (solid lines) whose non-emptiness indicates that we are unlikely to need any more complicated analysis. The dashed line is an eye-fitted line near the center of the bundle of piercing lines. Finally, the dotted lines, joining SCL corners of the split-diamond glyphs, indicates the range of possibilities for a true line (assuming that there is one).

exhibit 6

Birthrate against log size, where size (= (**) from exhibit 5) approximates the square root of the sum of the sizes of all places in a division

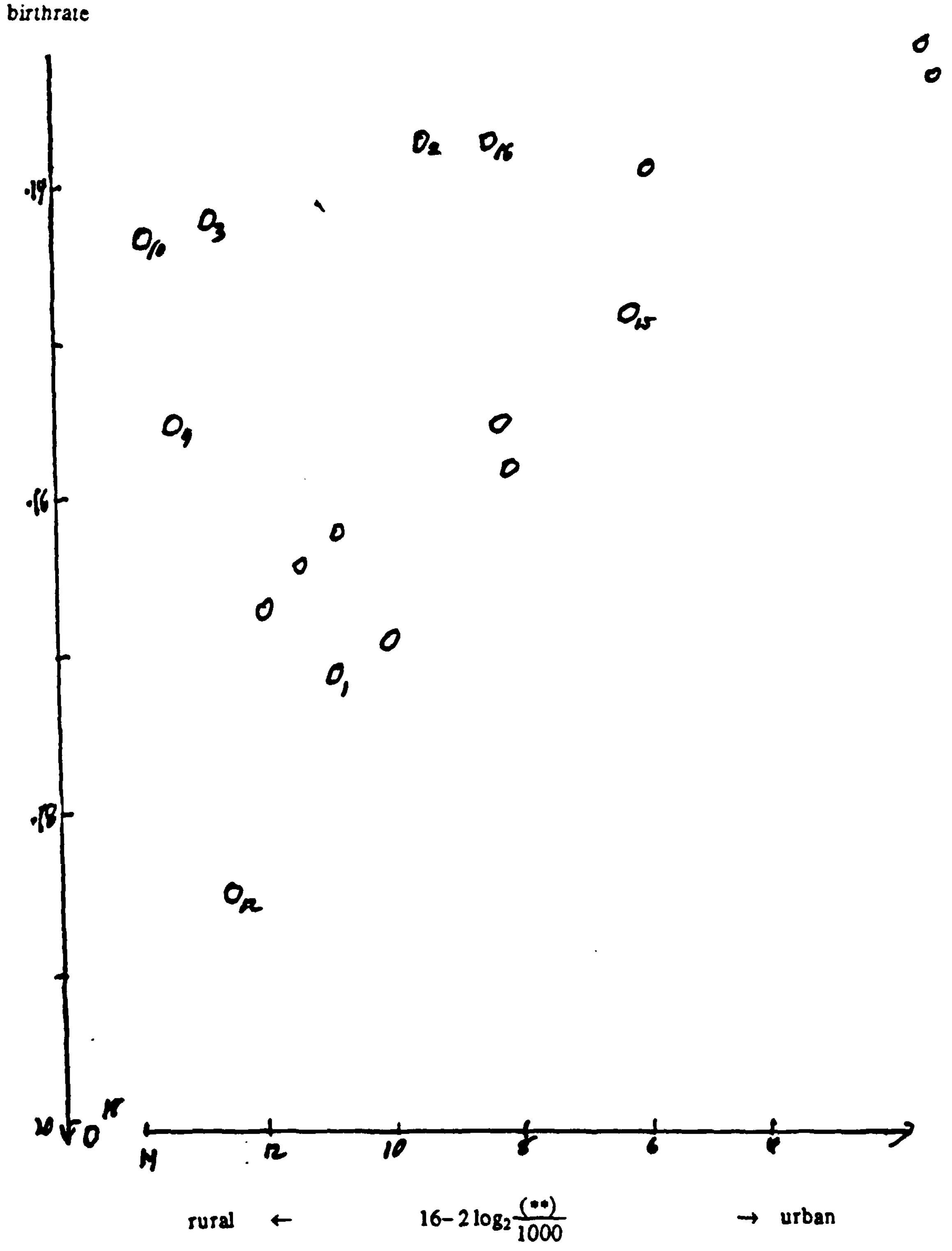


exhibit 7

As exhibit 6, but with squared horizontal scale

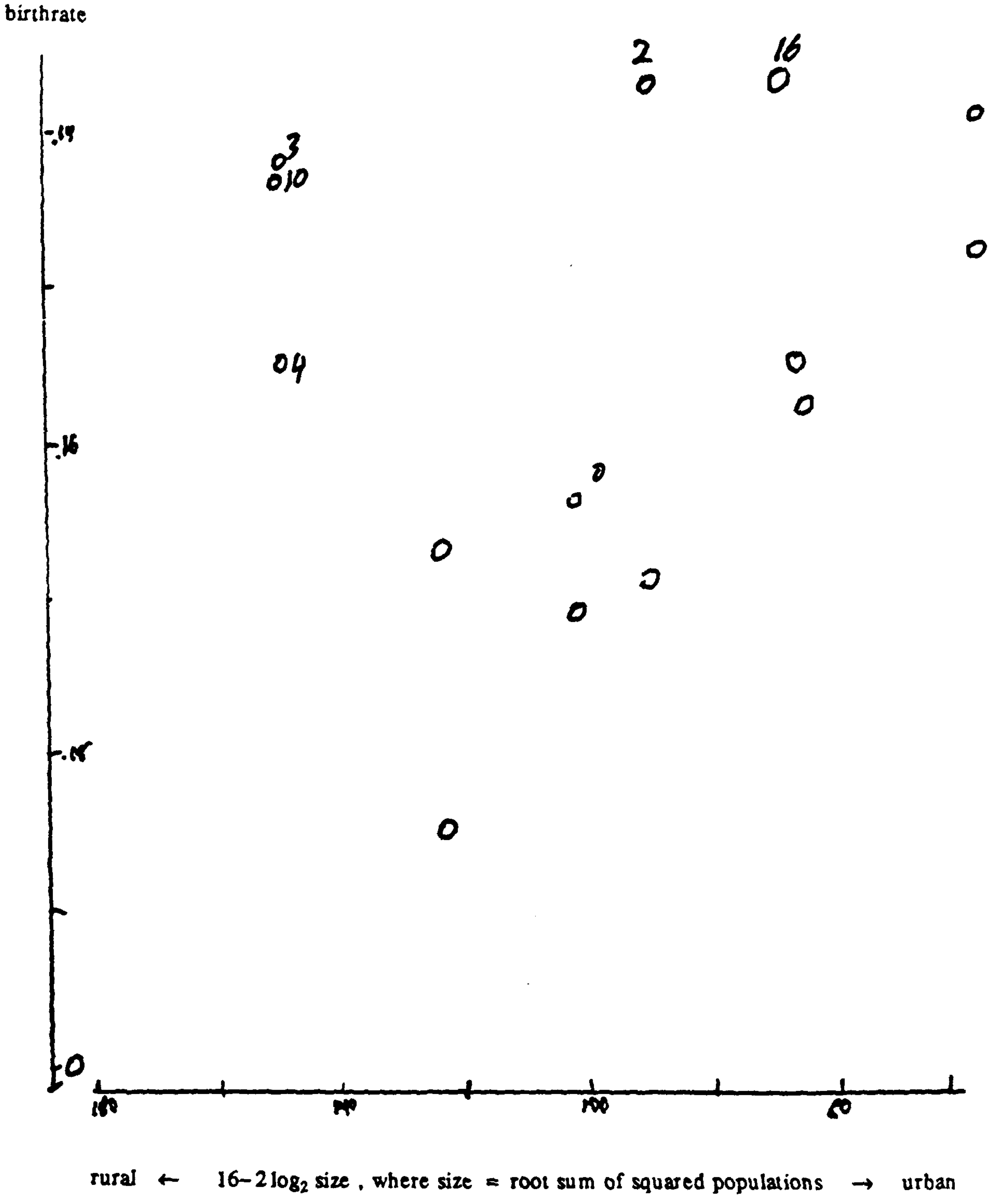


exhibit 9
about here

8. Further analysis

We can look harder at the data, in an overall, poorly informative way by summing the standardized squared deviations of the observed birth rates from some fitted lines (standardized to allow for the standard deviations that are consequences of assumed Poisson distribution). Doing this roughly, gives a sum of squares of 21.8, which is to be referred to 18 − 2 = 16 or 18 − 3 = 15 degrees of freedom. A reasonable threshold for choosing further analysis would probably be a ratio of 2 between sum of squares and degrees of freedom, which is clearly not even approached. Since exhibit 8 offers no specific indications suggesting further analysis, we are probably well advised to stop with our apparent linear dependence of birth rate on a simple measure of rurality.

It might be of interest to use actual populations of places under 1,000 and see what effect this would have on the analysis.

9. Kinds of consistency

We have now looked at the Saskatchewan data (birthrate vs. rural urban index) in two quite different ways - - both oriented toward: How well does the data fit a simple relation? Do we seem to need to look further? It is probably time that we compared these approaches in rather greater generality?

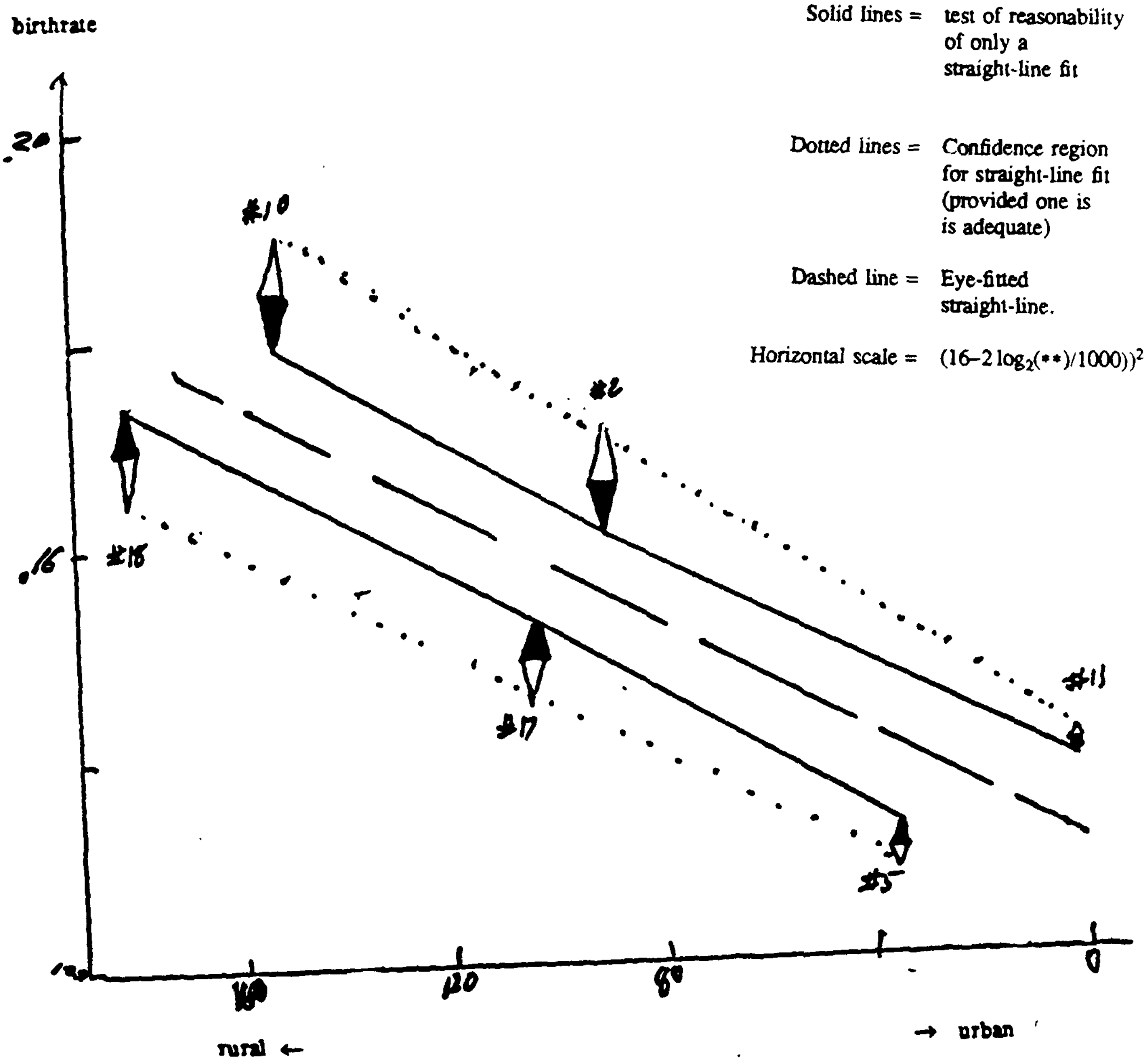
We could characterize the two approaches as one of small-group responsibility and another of collective violence. If a set of SQL-based apertures cannot be passed, there is at least one subset of 3 apertures that cannot be passed. Thus our feeling of inadequate fit can be assigned to one or more subsets of 3. (This is assigning responsibility to the smallest possible subset, since *any pair* of apertures (for different x-values) can be passed by each of many lines.)

To look at a sum of squared deviations, by contrast, is to blend all deviations into an unresolved whole. A positive result is a collective result! If we find a poor fit, but we are not allowed to look inside our omnibus statistic, we have no idea what it is that is mediating the poor fit. All we know is that, collectively, the deviations are too large - - that there is too much collective violence to the deviations.

Some will say that we should make our choice between two such approaches on the basis of power, but there are varied reasons why merely calling on the power concept does not work.

exhibit 9

Skeleton aperture plot for the Saskatchewan data



Solid lines = test of reasonability of only a straight-line fit

Dotted lines = Confidence region for straight-line fit (provided one is is adequate)

Dashed line = Eye-fitted straight-line.

Horizontal scale = $(16-2\log_2(**)/1000))^2$

First, we are in a multiresponse situation, and relatively high power in one direction is likely to correspond to relatively low power in another. We are forced to think about, for example, the least power for points on a hypersurface that encloses the null situation. Which hypersurface?

The situation is probably clearest when we are comparing several y's, with $\eta_i = \text{ave}\, y_i$, and no other quantities enter. If one looks at the hypersurface $\sum(\eta_i - \bar{\eta})^2$ = constant, the minimum power on that hypersurface is maximized by using the collective violence statistic, $\sum(y_i - \bar{y})^2$. If, on the other hand, one looks at the hypersurface range $\{\eta_i\}$ = constant, the minimum power on this hypersurface is maximized by the 2-value-responsibility statistic, range $\{y_i\}$. It is not enough to ask for "power", we must say where we want the power.

Second, the pure concept of power is inadequate to deal with collective violence. The idea of power grew up in the univariate situation - - usually a single comparison - - where the value was either "up so-and-so" or "down such-and-such", and where the sign of Student's t distinguished "up" from "down". I, long ago, introduced the notion of "useful power" as the product of mathematical (pure) power and the chance that if a definite answer were given, we would know what it meant. A moment's reflection shows that the useful power of any collective violence statistic is either zero, or very nearly zero. This is certainly the case for statistics based on $\sum(y_i - \bar{y})^2$. If we accept "useful power" as a reasonable concept - - a reasonable criterion - - then we will have to eschew collective violence statistics, and will probably find ourselves working with small-subset-responsibility statistics.

This is as true for "ave$\{y \mid x\}$ may be linearly dependent on x" as it is for "the η_i = ave$\{y_i\}$ may be all equal" or as it is for situations much more complex than either of these.

10. The case k=3

If k=3, and the values of x are equally spaced, we can write the values of y as

$$y_- = \mu_- + u_0/\sqrt{3} - u_1/\sqrt{2} \quad - u_2/\sqrt{6}$$

$$y_0 - \mu_0 + u_0/\sqrt{3} \qquad\qquad\quad - 2u_2/\sqrt{6}$$

$$y_+ = \mu_+ + u_0/\sqrt{3} + u_1/\sqrt{2} \quad + u_2/\sqrt{6}$$

with the null hypothesis represented by $u_{0,}\ u_{1,}\ u_2$ all uncorrelated Gaussians of equal variance, which we may as well take to be unit variance. The deviations of the y's from a straight line are specified by the terms in u_2. A line will just sneak through apertures of $\pm u_2 s_\nu$ (where ave$\{s_\nu^2\} = 1$) based on these y's when $|(u_2/\sqrt{6}) - (-2u_2/\sqrt{6})| > \pm(a-(-a))s_\nu$. The boundary case is

$$2a = \frac{3u_2}{\sqrt{6}} \cdot \frac{1}{s_\nu}$$

$$a = \frac{2}{2\sqrt{6}}(\frac{u_2}{s_\nu})$$

Since u_2/s_ν is distributed like Student's t on ν degrees of freedom, we get

$$a = \frac{3}{2\sqrt{6}} \cdot t_\nu[2.5\%] = .6124\, t_\nu[2.5\%]$$

with the result

ν/k	1	1.2	1.5	2	3	6	∞
SQL coefficient	1.948	(1.80)	(1.65)	1.498	1.385	1.287	1.200

Going to the tables in Technical Report 300, we see that

ν/k	1	1.5	2	3	6	∞
tight SPL coeff – SQL coeff	(1.01)	–	.67	.58	.51	.46
severe SCL – SPL	1.43	–	.97	.86	.75	.70
.68(severe SCL – severe SPL)	.98	–	.66	.58	.51	.48
tight SPL	2.96	–	2.17	1.97	1.80	1.66
diff of last two	1.98	–	1.51	1.39	1.29	1.18
tight SQL coeff (see above)	1.95	–	1.50	1.38	1.29	1.20

(The parenthetic values are both boldly extrapolated, so their disagreement can be neglected.)

As the last two lines show

$$\text{tight SQL} \doteq \text{tight SPL coeff} - .68(\text{severe SCL} - \text{severe SPL})$$

is a very good approximation. This relation has been developed for a very special case (k=3) of "tight" coefficients; brave people may wish to try it for general cases of tight coefficients.

11. An alternative approach

Another way to approximate SQL is to ask for what Q-value (left-hand area) would SPL equal the SQL for our standard tail area (Q = .95). A little inquiry into Harter - - leads to the following results

ν/k	1	2	3	6	∞
Q^*	86.3%	84.4%	82.6%	80.7%	79.5%

which is interesting, but not nearly as a simple approximation.

For the Saskatchewan example ($\nu = \infty$, k=18) the use of 79.3% would lead to an SQL of 4.22/2 = 2.11, not too far from the 2.16 found by the other extrapolation.

12. Appendix on severe SQL's

We now turn to the "severe" or "Bonferroni" approach. If we consider our "does a line pass through" problem carefully, we see that some line will pass all apertures if some line passes each set of 3 apertures. (We can see this inductively by starting with 3 apertures with smallest x's, and adding apertures one at a time from left to right. If there is difficulty at any step, the closest that a line passing all the previous apertures can come to passing the aperture being added will be determined by a line that contacts the edges of *two* of the previous apertures. Thus those two apertures, and the new aperture, make up a set of 3 that cannot be passed.)

For (x_a, y_a), (x_b, y_b) and (x_c, y_c) with $x_a \le x_b \le x_c$ the test statistic for passing is

$$y_b = \frac{(x_c - x_b)y_a + (x_b - x_a)y_c}{x_c - x_a}$$

whose variance is

$$\left[\frac{3}{2} + 2\left[\frac{x_b - x_a}{x_c - x_a} - 0.5\right]^2\right]\sigma^2 = \left[\frac{3}{2} + \frac{1}{2}t^2\right]\sigma^2 = R\sigma^2$$

There are

$$\binom{k}{3} = k(k-1)(k-2)/6$$

such triples, each of which can be too positive or too negative. Thus Bonferroni operates with $k(k-1(k-2)/3$ ends, and a $5\%/(k(k-1)(k-2)/3) = 15\%/k(k-1)(k-2)$ tail area for each.

The values of $R = 1.5 + .5t^2$ range from 3/2 to 2, where t varies from 0 to 1 and is likely to be somewhere near uniformly distributed. An approximate Bonferroni would use an averaged t^2. If we put 1/3 for our averaged t^2, $R = (3/2) + (1/2)(1/3) = 5/3 = 1.66$. (Some more detailed calculations suggest 1.64 may be a more precise value, but the difference is unimportant here.)

Thus an approximate Bonferroni calculation, usually slightly conservative because of the actual distribution of t and because of the nature of the averaging involved finds

$$\text{severe SQL's} \quad \text{at} \quad t_{\nu}[15\%/k(k-1)(k-2)]\,(\sqrt{1.66}\,s_{\nu})$$

Values thus obtained are given in exhibit 10. Notice that most entries have

severe SQL coefficient > severe SPL coefficient

although we know that

tight SQL coefficient < tight SPL coefficient.

exhibit 10

about here

The only reasonable conclusion is that trying to control the average number of triples which *cannot* be passed is too far away from controlling whether one or more triples cannot be passed for "severe" to be a reasonable choice.

This does not seem so surprising when we realize that, for $k=18$ (as in the Saskatchewan example) there are 18(17)16/6 = 816 triples generated by 18 apertures. Correlations of behavior of one triple with that of another must be substantial, and "failure to pass" must tend to occur, even in the null situation, for 2 or more triples at a time.

Thus we need to use tight SPL's or some close approximation thereto.

exhibit 10

Severe SQL coefficients
(calculated as $t_\nu\,[15\%/k\,(k-1)(k-2)]\,(\sqrt{1.66})$ see ** below)

k	$\nu/k=1$	$\nu/k=2$	$\nu/k=3$	$\nu/k=6$	$\nu/k=\infty$
3	2.050	1.576	1.457	1.307	1.262
4	2.780	2.065	1.890	1.740	1.609
5	3.074	2.307	2.116	1.952	1.808
6	3.190	2.452	2.261	2.094	1.947
8	3.297*	2.621	2.442	2.282	2.140
10	3.329*	2.721*	2.556*	2.408*	2.214
12	3.339*	2.790	2.639*	2.501*	2.375
15	3.343*	2.866	2.731*	2.606*	2.492*
18	3.345*	2.923	2.801*	2.688*	2.582*
20	3.347*	2.954*	2.840*	2.733*	2.02*
25	3.355*	3.019*	2.919*	2.825*	2.735*
30	3.366*	3.079*	2.982*	2.897*	2.816*
40	3.392*	3.152*	3.078*	3.007*	2.939*
50	3.419*	3.215*	3.151*	3.090*	3.303*
60	3.445*	3.266*	3.210*	3.156*	3.103*

*Larger than severe 95% SPL (!)

**Calculated using $\sqrt{1.66}$ factor - - maybe 1% too large

REFERENCES

Brillinger, D. R. (1990). "Two reports on the analysis of spatially aggregate data; a) Mapping aggregate birth data, (b) Spatial-temporal modeling of spatially aggregate birth data," *Technical Report No. 244,* March 1990, Department of Statistics, University of Califomia, Berkeley.

Hoaglin, D. C. and Tukey, J. W. (1985n). "When is a point discrepant"?, Section 9C of *Exploring Data Tables, Trends and Shapes,* eds. Hoaglin, Mosteller, Tukey. John Wiley, ew York.

Kafadar, K. and Tukey, J. W. (1991). An approach to U. S. Cancer death rates involving urbanness and geographic contiguity: 1: A simple adjustment for urbanness, submitted to *International Statistical Review.*

Tukey, J. W. (1977a). *Exploratory Data Analysis,* First Edition, Addison-Wesley, Reading, MA.

Tukey, J. W. (1990). "Combining CL (confidence limits) and PL (partial limits) displays," *Technical Report No. 300,* Department of Statistics, Princeton University, Princeton, NJ 08544-1000.

NOTE: Letters used with years on papers correspond to bibliography in all volumes of the *Collected Works of John W. Tukey.*